BAYES CLASSIFICATION AND SELECTION OF ALLOCATION VARIABLES WHEN SAMPLING UNITS ARE SUBSAMPLED

J. C. Evans
N.S.W. Department of Agriculture
Haymarket 2000
Australia

D. S. Robson and S. J. Schwager
Cornell University
Ithaca, NY 14853
U.S.A.

BU-872-M September 1985

Abstract. Multinormal Bayes rules are given for classifying a subsampled sampling unit by its mean vector, including the case when the number of subsamples is proportional to the size of the unit. The allocation variables selection methods of Evans *et al.* (1985a,b) are extended to the subsampling situation. An example of the implementation of these rules and methods is given.

Key Words and Phrases. Bayes classification, subsampling, allocation variables selection, remote sensing, estimated Bayes risk.

Running Title. Bayes Classification of Subsampled Sampling Units

1. Introduction

Evans et al. (1985a) summarized the Bayes rule for classifying a sampling unit on the basis of a single multivariate observation. But in practice it may be necessary to classify a sampling unit on the basis of the mean vector of several multivariate observations arising from its subsampling. Except in the special case of a constant member of subsamples per unit, no Bayes rule exists for this purpose. A rule is required that takes into account unequal numbers of subsamples per sampling unit. Also, as in the case of no subsampling, methods are needed for selecting a subset of variables that maximizes classification accuracy and reduces computational costs.

Section 2 presents Evans' (1984) multivariate normal Bayes rules for classifying subsampled units in the cases of the number of subsamples being proportional or not proportional to the size of the sampling unit. In the case of subsampling proportionally to size, rules are given for both when misclassification costs are also proportional to size and when they are not. Section 3 generalizes to the cases of subsampling the all-possible-subsets algorithm of Evans et al. (1985a), for isolating subsets of variables that minimize estimated Bayes risk, and the stagewise algorithms of Evans et al. (1985b) that use tests of additional reduction in Bayes risk to select a subset for allocation purposes. Section 4 gives a simple example to illustrate the theory of Sections 2 and 3.

2. Bayes Classification of Subsampled Sampling Units

Denote the multivariate observation on subsampling unit $k \in \{1,\cdots,n\}$ of an unidentified sampling unit by $\underset{\sim}{y}_k \equiv \left(\underset{\sim}{y}'_{ik},\cdots,\underset{\sim}{y}'_{Tk}\right)'$, where $\underset{\sim}{y}_{gk}$ corresponds to a group $g \in \{1,\cdots,T\}$ of $r \geq 1$ variables. Multinomial

Bayes rules will here be developed for classifying a sampling unit on the basis of its mean $\underline{\bar{y}}_{\cdot} \equiv \sum_{k=1}^{n} \underline{y}_k/n$, where the n subsamples have been obtained independently, as being from class $i' \in \{1,\cdots,k\}$. Before any unidentified sampling units can be classified, usually the parameters of a classification rule first need to be estimated from observations on $N \equiv \sum_{i=1}^{K} r_i$ independent reference sampling units and then the rule tested on $M = \sum_{i=1}^{K} m_i$ independent cross-validation sampling units. Denote the observation vector on subsampling unit $k \in \{1,2,\cdots,n_{ij}\}$ of sampling unit $j \in \{1,\cdots,r_i,\cdots, r_i+m_i\}$ from class $i \in \{1,\cdots,K\}$ by $\underline{y}_{ijk} \equiv \left(\underline{y}'_{1ijk},\cdots,\underline{y}_{Tijk}\right)'$.

Assume that the model

$$\underline{y}_{ijk} = \underline{\mu}_i + \underline{\varepsilon}_{ij} + \underline{\delta}_{ijk} \tag{1}$$

is appropriate, where $\underline{\mu}_i$ is fixed but unknown, $\underline{\varepsilon}_{ij} \sim$ i.i.d. $N\left(\underline{0},\Sigma_{\varepsilon_i}\right)$, $\delta_{ijk} \sim N\left(\underline{0},\Sigma_{\delta_i}\right)$, and the ε_{ij} terms are independent of the δ_{ijk} terms. Then it follows that $E(\underline{y}_{ijk}) = \underline{\mu}_i$, $V(\underline{y}_{ijk}) = \Sigma_{\varepsilon_i} + \Sigma_{\delta_i}$, $\mathrm{Cov}(\underline{y}_{ijk},\underline{y}_{ijk'}) = \Sigma_{\varepsilon_i}$ for $k \neq k' = 1,\cdots,n_{ij}$, and $\mathrm{Cov}(\underline{y}_{ij'k'},\underline{y}_{ijk}) = 0$ for $j \neq j' = 1,\cdots,r_i + m_i$. Although other models and conditions may be more appropriate for specific problems, only the model of Equation (1) will be used here.

Under this model, for a sampling unit $j \in \{1,\cdots,r_i+m_i\}$ from class $i \in \{1,\cdots,K\}$, the variance of $\underline{\bar{y}}_{ij\cdot} \equiv \sum_{k=1}^{n_{ij}} \underline{y}_{ijk}/n_{ij}$ for a fixed n_{ij} is

$$V\left(\underline{\bar{y}}_{ij\cdot}\,|n_{ij}\right) = \Sigma_{\varepsilon_i} + \Sigma_{\delta_i}/n_{ij} \equiv \Sigma_i \quad . \tag{2}$$

Thus, if an unidentified sampling unit has come from class i, $E(\underline{\bar{y}}_{\cdot}|n) = \underline{\mu}_i$ and $V(\underline{\bar{y}}_{\cdot}|n) = \Sigma_i = \Sigma_{\varepsilon_i} + \Sigma_{\delta_i}/n$, and so that multivariate normal probability density function (p.d.f.) is given by

$$f_i(\underline{\bar{y}}_{\cdot}|n) \equiv (2\pi)^{-rT/2}\left|\Sigma_i\right|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(\underline{\bar{y}}_{\cdot} - \underline{\mu}_i)'\,\Sigma_i^{-1}\,(\underline{\bar{y}}_{\cdot} - \underline{\mu}_{\cdot})\right\} \quad . \tag{3}$$

Unlike the case of no subsampling, Σ_i has to be found for each different n specified. But usually $\underline{\mu}_i$, Σ_{ε_i}, and Σ_{δ_i}, and thus Σ_i, are unknown and need to be estimated from the reference observations that are available on fixed but unequal numbers of subsamples per sampling unit. Unbiased estimators of $\underline{\mu}_i$ and Σ_{δ_i} are then given, respectively, by

$$\bar{\underline{y}}_{i..} = \sum_{j=1}^{r_i} \sum_{k=1}^{n_{ij}} \underline{y}_{ijk}/n_{i.}, \quad n_{i.} = \sum_{j=1}^{r_i} n_{ij} \quad , \tag{4}$$

and

$$S_{\delta_i} = W_{\delta_i} / (n_{i.}-\tau_i) \quad , \tag{5}$$

where

$$W_{\delta_i} \equiv \sum_{j=1}^{r_i} \sum_{k=1}^{n_{ij}} \left(\underline{y}_{ijk}-\bar{\underline{y}}_{ij.}\right)\left(\underline{y}_{ijk}-\bar{\underline{y}}_{ij.}\right)'$$

is the among-subsamples sums of squares and products (SSP) matrix for class i. Evans (1984) found an unbiased estimator of Σ_{ε_i} to be

$$S_{\varepsilon_i} = \left(S_i-S_{\delta_i}\right)\left(r_{i-1}\right) / \left(n_{i.} - \sum_{j=1}^{r_i} n_{ij}^2/n_{i.} \right) \quad , \tag{6}$$

where

$$S_i = W_i/(r_{i-1}) \quad \text{and} \quad W_i = \sum_{j=1}^{r_i} n_{ij}\left(\bar{\underline{y}}_{ij.}-\bar{\underline{y}}_{i..}\right)\left(\bar{\underline{y}}_{ij.}-\bar{\underline{y}}_{i..}\right)' \tag{7}$$

is the among-samples SSP matrix for class i. An unbiased estimator of $\Sigma_i = \Sigma_{\varepsilon_i} = \Sigma_{\delta_i}/n$ is then given by

$$\hat{\Sigma}_i = S_{\varepsilon_i} + S_{\delta_i}/n \quad . \tag{8}$$

In the special case of a constant number n of subsamples for all sampling units, Σ_{ε_i} and Σ_{δ_i} need not be estimated, as the single Σ_i can be estimated directly and unbiasedly by

$$\hat{\Sigma}_i = S_i/n \quad . \tag{9}$$

Only in the case of constant n do the estimators $\bar{\underline{y}}_{i..}$, S_{δ_i}, and $\hat{\Sigma}_i$ have good properties apart from being unbiased. Those given here are generalizations from the univariate case (Arnold, 1981, Lemma 15.5 and Theorems 15.3, 15.6, and 15.7) to the multivariate case, for the balanced one-way random effects model which is appropriate for each class. Jointly, these estimators constitute a complete sufficient statistic for this multivariate balanced one-way random effects model. Individually, they are minimum generalized-variance unbiased estimators of $\underline{\mu}_i$, Σ_{δ_i}, and Σ_i. (This follows directly from their joint complete sufficiency after applying known theory as given in Mood, Graybill and Boes, 1974, pp. 351-356.) By also using Theorem 18.5 of Arnold (*loc. cit.*), it follows that maximum likelihood estimators of $\underline{\mu}_i$, Σ_i, and Σ_{ε_i}, when $rT \leq r_i - 1$, are $\bar{\underline{y}}_{i..}$,

$$\hat{\Sigma}_i^* \equiv (r_i-1)S_i/r_i n = (r_i-1)\,\hat{\Sigma}_i/r_i \quad , \tag{10}$$

and

$$\hat{\Sigma}_{\varepsilon_i} = \begin{cases} \hat{\Sigma}_i^* - S_{\delta_i}/n & \text{if nonnegative definite} \\ 0 & \text{otherwise} \quad . \end{cases} \tag{11}$$

An unbiased estimator of Σ_{ε_i} is obtained from Equation (6) by substituting n for each n_{ij} to give

$$S_{\varepsilon_i} = \hat{\Sigma}_i - S_{\delta_i}/n \quad . \tag{12}$$

But it follows from Theorem 15.6 of Arnold (*loc. cit.*) that neither this nor any other unbiased estimator of Σ_{ε_i} is nonnegative definite with probability 1, although S_{ε_i} could be substituted for $\hat{\Sigma}_i^* - S_{\delta_i}/n$ in Equation (11) to give another biased nonnegative definite estimator. The choice among these estimators of Σ_{ε_i} need not be made here, as none is required to apply the Bayes rule of Subsection 2.1 for this special case of

constant n for all sampling units. Similarly, as no estimator of Σ_{δ_i} is required, no choice needs to be made here between S_{δ_i} and other estimators such as a multivariate analogue to the maximum likelihood estimator given by Arnold (*loc. cit.*) for the univariate case. According to the properties ascribed to $\bar{\underline{y}}_{i..}$ as an estimator of $\underline{\mu}_i$, it is the appropriate estimator here. Of the two estimators of Σ_i given above, the unbiased estimator $\hat{\Sigma}_i$ of Equation (9) is used subsequently rather than the maximum likelihood estimator $\hat{\Sigma}_i^*$ of Equation (10).

Unfortunately, in the case of different n_{ij} among reference sampling units, unbiasedness is all that can be said in favour of the estimators $\bar{\underline{y}}_{i..}$, S_{δ_i}, and S_{ε_i} in Equations (4), (5), and (6), and thus of $\hat{\Sigma}_i$ in Equation (8). With regard to the unbalanced one-way random effects model, which is appropriate here for each class $i \in \{1,\cdots,K\}$, Arnold (1981, p. 269) states for the univariate case:

> "The class of random effects ... models for which optimal procedures ... exist seems to be a very small class. We have already seen that it is not possible to find ... nonnegative unbiased estimators for even such simple random effects models as the balanced one-way ... models ... . For unbalanced models the situation is even worse. Typically, there is no complete sufficient statistic for the model, and no minimum variance unbiased estimators. ... However, they (these models) do arise in practice, and many procedures for analyzing them have been suggested (see Searle, 1971, pp. 376-514 and Harville, 1977 for a description of some of these procedures)."

Obviously, at least the same problems exist in the multivariate case. Nevertheless, $\bar{\underline{y}}_{i..}$ and S_{δ_i} will be used here as estimators of $\underline{\mu}_i$ and Σ_{δ_i}. But S_{ε_i} of Equation (6) will be replaced by a nonnegative estimator

$$\hat{\Sigma}_{\varepsilon_i} \equiv \begin{cases} S_{\varepsilon_i} & \text{if nonnegative definite} \\ 0 & \text{otherwise} \end{cases} . \tag{13}$$

Then an estimator of Σ_i is obtained by replacing S_{ε_i} by $\hat{\Sigma}_{\varepsilon_i}$ in Equation (8) to give

$$\hat{\Sigma}_i = \hat{\Sigma}_{\varepsilon_i} + S_{\delta_i}/n \quad . \tag{14}$$

When it is assumed that $\Sigma_{\delta_1} = \cdots = \Sigma_{\delta_K} \equiv \Sigma_\delta$, $\Sigma_{\varepsilon_1} = \cdots = \Sigma_{\varepsilon_K} \equiv \Sigma_\varepsilon$, and thus $\Sigma_1 \cdots = \Sigma_K \equiv \Sigma$ for a given n, as is done in the example of Section 4, an unbiased estimator of $\Sigma = \Sigma_\varepsilon + \Sigma_\delta/n$ is given by

$$\hat{\Sigma} = S_\varepsilon + S_\delta/n \quad , \tag{15}$$

where

$$S_\delta = \sum_{i=1}^{K} (n_{i\cdot} - r_i) S_{\delta_i} \Big/ \sum_{i=1}^{K} (n_{i\cdot} - r_i) \quad , \tag{16}$$

and

$$S_\varepsilon = (S - S_\delta)(N-K) \Big/ \left(n_{\cdot\cdot} - \sum_{i=1}^{K} \frac{1}{n_{i\cdot}} \sum_{j=1}^{r_i} n_{ij}^2 \right) \tag{17}$$

are unbiased estimators of Σ_δ and Σ_ε, respectively,

$$S = \sum_{i=1}^{K} (r_i - 1) S_i \Big/ \sum_{i=1}^{K} (r_i - 1) \quad \text{and} \quad n_{\cdot\cdot} = \sum_{i=1}^{K} \sum_{j=1}^{r_i} n_{ij} \quad . \tag{18}$$

To obtain a nonnegative definite estimator of Σ_ε, replace S_e by

$$\hat{\Sigma}_\varepsilon = \begin{cases} S_\varepsilon & \text{if nonnegative definite} \\ 0 & \text{otherwise} \quad . \end{cases} \tag{19}$$

and then a new estimator of Σ is given by

$$\hat{\Sigma} = \hat{\Sigma}_\varepsilon + S_\delta/n \quad . \tag{20}$$

When n is also constant, Equations (15) – (17) reduce to

$$\hat{\Sigma} = S/n \quad , \tag{21}$$

$$S_\delta = \sum_{i=1}^{K} r_i S_{\delta_i} / N \quad , \tag{22}$$

and

$$S_\varepsilon = \hat{\Sigma} - S_\delta/n \quad . \tag{23}$$

The estimators $\hat{\Sigma}$ and S_δ then have some good properties. Earlier, it was stated that $\bar{y}_{i..}$, $\hat{\Sigma}_i$, and S_{δ_i} are minimum generalized-variance unbiased estimators and also comprise a complete sufficient statistic for the multivariate single-class balanced one-way random effects model. It follows that $\left(\bar{y}_{i..}, i=1,\cdots,K; \hat{\Sigma}; S_\delta\right)$ is a complete sufficient statistic for the K-classes extension of this model with common Σ and common Σ_δ, and that each component is a minimum generalized-variance unbiased estimator. By also using Theorems 18.18, 19.1, and 19.2 of Arnold (1981), analogues of the earlier Equations (10) and (11) here give maximum likelihood estimators of μ_i, Σ, and Σ_ε, when $rT \leq N-K$, as $\bar{y}_{i..}$, $i=1,\cdots,K$,

$$\hat{\Sigma}^* \equiv (N-K)\,\hat{\Sigma}/N \quad , \tag{24}$$

and

$$\hat{\Sigma}_\varepsilon \equiv \begin{cases} \hat{\Sigma}^* - S_\delta/n & \text{if nonnegative definite} \\ 0 & \text{otherwise} \quad . \end{cases} \tag{25}$$

The unbiased estimator $\hat{\Sigma}$ of Equation (21) will be used subsequently here. But in the earlier case of unequal n_{ij}, the nonnegative definite estimators $\hat{\Sigma}_\varepsilon$ and $\hat{\Sigma}$ of Equations (19) and (20), and the unbiased estimator S_δ of Equation (16), are subsequently used.

2.1 Fixed Unequal Numbers of Subsamples Not Proportional to Sizes of Sampling Units

A multivariate normal Bayes rule is given here for classifying the cross-validation sampling units $j=1,\cdots,m_i$ from each class $i=1,\cdots,K$ based on the vectors of means $\bar{y}_{ij.}$ of n_{ij} subsamples, and for classifying an unidentified sampling unit based on the vector of means $\bar{y}_{.}$ of an arbitrary n subsample. It is assumed that, as for reference sampling units, the numbers of subsamples n_{ij} or n were fixed nonproportionately to the sizes of the sampling units.

The multinormal p.d.f., $f_i(\bar{\underline{y}}_{\cdot}|n)$, was given in Equation (3). It is estimated as $\hat{f}_i(\bar{\underline{y}}_{\cdot}|n)$ by replacing $\Sigma_i = \Sigma_{\varepsilon_i} + \Sigma_{\delta_i}/n$ by $\hat{\Sigma}_i$ of either Equations (14) or (9), whichever is appropriate, or in the case of a common $\Sigma = \Sigma_{\varepsilon} + \Sigma_{\delta}/n$, by $\hat{\Sigma}$ of either Equation (20) or (21), and $\underline{\mu}_i$ by $\bar{\underline{y}}_{i\cdot\cdot}$. It is required that $rT \leq r_i - 1$ when using $\hat{\Sigma}_i$ or $rT \leq N-K$ if using $\hat{\Sigma}$.

The Bayes classification rule, directly analogous to the no-subsampling Bayes rule of Evans et al. (1985a), is to classify a mean observation $\bar{\underline{y}}_{\cdot}$ based on an arbitrary n subsample as coming from the class $i' \in \{1,\cdots,K\}$ that minimizes $\sum_{\substack{i=1 \\ i\neq i'}}^{K} \pi_i C(i'|i) f_i(\bar{\underline{y}}_{\cdot}|n)$, where π_i is the prior probability that the sampling unit came from class i and $C(i'|i)$ is the cost of misclassifying a sampling unit from class i as being from class i'. This rule minimizes the Bayes risk

$$R = \sum_{i=1}^{K} \pi_i \sum_{\substack{i'=1 \\ i'\neq i}}^{K} C(ii'|i)P(i'|i) \quad , \tag{26}$$

where $P(i'|i)$ is the probability of misclassifying, on the basis of the variable $\bar{\underline{y}}_{\cdot}$, a sampling unit from class i as i'. When parameters have been replaced by estimators as above, $\hat{f}_i(\bar{\underline{y}}_{\cdot}|n)$ replaces $f_i(\bar{\underline{y}}_{\cdot}|n)$ in the classification rule.

2.2. Variable Numbers of Subsamples Proportional to Sizes of Sampling Units

In this subsection, although fixed n_{ij} are still assumed for reference sampling units, the n_{ij} and n for other units are proportional to their sizes. In the latter case, to signify that n_{ij} and n denote the numbers of subsamples determined by proportionality to sizes of cross-validation and unidentified sampling units, they are replaced here by x_{ij} and x as observations of a random variable X. As size is usually a random variable, and as its distribution may differ from class to class, X is considered as an additional allocation variable with a p.d.f., $f_i(x)$, for classes $i=1,\cdots,K$. Evans (1984) considered the normal and lognormal p.d.f.'s as examples of $f_i(x)$. When the parameters of $f_i(x)$ are unknown and need to be estimated to give $\hat{f}_i(x)$, size must also be measured on the reference sampling units $j=1,\cdots,r_i$ of class i that have been subsampled nonproportionately to size. Then it is known what the numbers of subsamples would have been, given the proportionality-to-size constant used for test and unidentified units, if the numbers of subsamples had been fixed proportionately to size.

2.2.1. Misclassification Costs Not Proportional to Size

Replace the multinormal p.d.f., $f_i(\underline{\bar{y}}_.)$, of Equation (3) by

$$f_i(x,\underline{\bar{y}}_.) = f_i(x)f_i(\underline{\bar{y}}_.|x) \quad , \tag{17}$$

where $f_i(\underline{\bar{y}}_.|x)$ is the same as Equation (3) under the assumption that $\sum_{\varepsilon_i}$ and $\sum_{\delta_i}/x$ still holds. This p.d.f. is estimated by

$$\hat{f}_i(x,\underline{\bar{y}}_.) = \hat{f}_i(x)\hat{f}_i(\underline{\bar{y}}_.|x) \quad , \tag{28}$$

which is then used in place of $_i(\underline{\bar{y}}_.|n)$ in the Bayes rule of Section 2.1 to give a classification based on both varieties X and $\underline{\bar{Y}}_.$. Evans (1984)

gave standard estimators for obtaining $\hat{f}_i(x)$ in the normal and lognormal cases. Estimators used to find $\hat{f}_i(\bar{y}_\cdot|x)$ are as in Section 2.1 above.

2.2.2. Misclassification Costs Proportional to Size

In addition to the number of subsamples x being proportional to size, the cost of misclassification may also be, that is,

$$C(i',x|i) \equiv xC(i'|i) \quad . \tag{29}$$

From first principles, Evans (1984) derived the Bayes classification rule for this situation, as now summarized in its simplest form. For each $i'=1,\cdots,K$, define a variable $Q_{i'}$ whose possible values for an observation $(x,\bar{y}_\cdot)$ are given by

$$q_{i'} \equiv \begin{cases} x \text{ if } i' \text{ minimizes } \sum_{\substack{i=1 \\ i\neq i'}}^{K} \pi_i C(i'|i) f_i(x,\bar{y}_\cdot) \\ 0 \qquad \text{otherwise} \end{cases} , \tag{30}$$

but with the restriction in the case of a nonunique minimum that only one of $q_1,\cdots,q_K$ is x. Then classify the sampling unit into the class i' for which $q_{i'} = x$. This rule minimizes the Bayes risk

$$R = \sum_{i=1}^{K} \pi_i \sum_{\substack{i'=1 \\ i'\neq i}}^{K} C(i'|i) E_i(Q_{i'}) \quad , \tag{31}$$

where E_i denotes expectation for class i. By inspection of Equation (30), it can be seen that the presence of a factor x (or any positive increasing function of x in its place) in the terms to be minimized does not change the choice of i' from that of Subsection 2.2.1. But it does change the Bayes risk itself, with $E_i(Q_{i'})$ here replacing P(i'|i) there, where P(i'|i) is the probability of misclassifying, on the basis of the variables X and

$\bar{\underline{Y}}_{i\cdot}$, a unit from class i as being from class i'. This fact is important later when choosing a subset that minimizes estimated Bayes risk. When $f_i(x,\bar{\underline{y}}_{\cdot})$ is unknown it is replaced in Equation (30) by $\hat{f}_i(x,\bar{\underline{y}}_{\cdot})$, which is defined as in Subsection 2.2.1.

3. Stagewise Algorithms for Selecting Groups of Allocation Variables

Evans et al. (1985a) proposed standardized estimated Bayes risk as a criterion for comparing subsets in an all-possible-subsets algorithm. Evans et al. (1985b) gave a test of additional reduction in Bayes risk for use in selecting a subset by stepwise, "minimal-best", or stepdown algorithms. All of these methods were for the case of no subsampling but will be readily generalized here to the cases of subsampling.

3.1. Fixed Unequal Numbers of Subsamples Not Proportional to Sizes of Sampling Units

Based on the v groups of variables, $\bar{\underline{Y}}_{g\ell\cdot}$, $\ell=1,\cdots,v\leq T$, constituting $\bar{\underline{Y}}_{\cdot}^{(v)} \equiv \left(\bar{\underline{Y}}'_{g_i\cdot},\cdots,\bar{\underline{Y}}'_{g_v}\right)'$, the Bayes rule of Section 2.1 can be used to classify the test sampling units $j=1,\cdots,m_i$ from each class $i=1,\cdots,K$ based on their vectors of means $\bar{\underline{y}}_{ij\cdot}^{(v)}$ over n_{ij} subsamples. Several notational changes are required. The p.d.f. $f_i(\bar{\underline{y}}_{\cdot}^{(v)}|n)$ is estimated as $f_i(\bar{\underline{y}}_{\cdot}^{(v)}|n)$ by replacing $\Sigma_i^{(v)} = \Sigma_{\varepsilon_i}^{(v)} + \Sigma_{\delta_i}^{(v)}/n$ by $\hat{\Sigma}_i^{(v)}$, or in the case of a common $\Sigma^{(v)} = \Sigma_{\varepsilon}^{(v)} + \Sigma_{\delta}^{(v)}/n$, by $\hat{\Sigma}^{(v)}$, and $\underline{\mu}_i^{(v)}$ by $\bar{\underline{y}}_{i\cdot\cdot}^{(v)}$. Then $\hat{f}_i(\bar{\underline{y}}_{\cdot}^{(v)}|n)$ is used in place of $f_i(\bar{\underline{y}}_{\cdot}|n)$ of the Bayes rule of Section 2.1. The Bayes risk formula of Equation (26) is replaced by one based on groups $g_1,\cdots,g_v$, namely

$$R_v = R(g_1,\cdots,g_v) = \sum_{i=1}^{K} \pi_i \sum_{\substack{i'=1 \\ i'\neq i}}^{K} C(i'|i) P_v(i'|i) \qquad (32)$$

where $P_v(i'|i)$ is the probability of misclassifying a sampling unit from

class i as being from class i' on the basis of $\underline{Y}_{\cdot}^{(v)}$.

Let $Q_{i'\cdot i}^{(v)}$ here be a variable that takes either value 1 if a sampling unit from class i is classified into class i' or 0 otherwise. Observed values of $Q_{i'\cdot i}^{(v)}$ on the test units $j=1,\cdots,m_i$ are denoted by $q_{i'ij}^{(v)}$. Then each $P_v(i'|i)$ is estimated by the proportion of class i test units misclassified as being from class i', namely $P_v(i'|i) = \bar{q}_{i'i\cdot}^{(v)}$, and used in place of $P_v(i'|i)$ in Equation (32) to give estimated Bayes risk, $\hat{R}_v$. Evans et al. (1985a) derived an estimate $\hat{V}(\hat{R}_v)$ of the variance of $\hat{R}_v$ to define the standardized estimated Bayes risk

$$z_v = \hat{R}_v / [\hat{V}(\hat{R}_v)]^{\frac{1}{2}} \quad . \tag{33}$$

They then used z_v as the criterion for finding the best subset of each size in an all-possible-subsets algorithm; that is, to find the subset of groups $g_1,\cdots,g_v$ that minimizes $z_v = z(g_1,\cdots,g_v)$ for each size $v=1,2,\cdots$. Exactly the same procedure can now be applied here.

Evans et al. (1985b) gave

$$\begin{aligned} \hat{R}_{u\cdot v} &= \hat{R}(g_{v+1},\cdots,g_{v+u}|g_1,\cdots,g_v) \\ &= \hat{R}_v - \hat{R}_{u+v} \end{aligned} \tag{34}$$

as an estimator of the decrease in Bayes risk, $R_{u\cdot v}$, due to adding u groups to the current v groups, and defined and gave $\hat{V}(\hat{R}_{u\cdot v})$ as an estimator of its variance. Then the standardized decrease in estimated Bayes risk,

$$z_{u\cdot v} = \hat{R}_{u\cdot v} / [\hat{V}(\hat{R}_{u\cdot v})]^{\frac{1}{2}} \quad , \tag{35}$$

was used in a one-sided test of $H_0: R_{u+v} = R_v$ versus $H_1: R_{u+v} < R_v$ by comparing $z_{u\cdot v}$ with the upper $100(1-\alpha)\%$ point Z_α of the N(0,1) distribution. This test was used at each step of stepwise, simultaneous stepdown,

and minimal-best classification algorithms to select a subset of groups. The minimal-best algorithm finds the smallest of the best subsets, from the all-possible-subsets classifications, that retains most of the allocation accuracy of the full set of groups. Exactly the same test and algorithms are applicable here.

3.2. Variable Numbers of Subsamples Proportional to Sizes of Sampling Units — Misclassification Costs Not Proportional to Sizes

Based on $\bar{\underline{y}}_{\cdot}^{(v)}$ and X, the Bayes rule of Section 2.2.1 can be used to classify the test sampling units $j=1,\cdots,m_i$ from class $i=1,\cdots,K$ based on their $\bar{\underline{y}}_{ij\cdot}^{(v)}$ and x_{ij} values. The p.d.f. $f_i(x,\bar{\underline{y}}_{\cdot}^{(v)}|x)$ is estimated as $\hat{f}_i(x,\bar{\underline{y}}_{\cdot}^{v} = \hat{f}_i(x)\hat{f}_i(\bar{\underline{y}}_{\cdot}^{(v)}|x)$, where $\hat{f}_i(\bar{\underline{y}}_{\cdot}^{(v)}|x)$ is defined as in Section 3.1. Then $\hat{f}_i(x,\bar{\underline{y}}_{\cdot}^{(v)})$ is used in place of $\hat{f}_i(x,\bar{\underline{y}}_{\cdot})$ of Section 2.2.1. The Bayes risk formula of Equation (32) of Section 3.1 is replaced here by

$$R_{X,v} = R(X,g_1,\cdots,g_v) \tag{36}$$

with $P_v(i'|i)$ replaced by $P_{X,v})i'|i)$. To define $\hat{R}_{X,v}$, other notational changes are needed from Section 3.1 to here, namely replacing $Q_{i'\cdot i}^{(v)}$, $q_{i'ij}^{(v)}$ and $p_v(i'|i) = q_{i'i\cdot}^{-(v)}$ by $Q_{i'\cdot i}^{(X,v)}$, $q_{i'ij}^{(X,v)}$ and $p_{X,v}(i'|i) = q_{i'i\cdot}^{-(X,v)}$, respectively. Then $p_{X,v}(i'|i)$ is used in place of $P_{X,v}(i'|i)$ in Equation (36) to give $\hat{R}_{X,v}$ which is standardized exactly as in Section 3.1 to give

$$z_{X,v} = \hat{R}_{X,v}/[\hat{V}(\hat{R}_{X,v})]^{\frac{1}{2}} \quad . \tag{37}$$

Now $z_{X,v}$ can be used in an all-possible-subsets algorithm to find the last subset of groups of each size $v=1,2,\cdots$, in the presence of X, namely the groups $g_1,\cdots,g_v$ that minimize $z_{X,v} = z(X,g_1,\cdots,g_v)$.

An estimator of the decrease in Bayes risk due to adding u groups to the current v groups in the presence of X is given by

$$\hat{R}_{u\cdot X,v} = \hat{R}(g_{v+1},\cdots,g_{v+u}|X,g_1,\cdots,g_v)$$
$$= \hat{R}_{X,v} - \hat{R}_{X,u+v} \quad , \tag{38}$$

whose estimated variance $\hat{V}(\hat{R}_{u\cdot X,v})$ is found in the same way as $\hat{V}(\hat{R}_{u\cdot v})$ in Section 3.1 after making the notation changes here. Now

$$z_{u\cdot X,v} = \hat{R}_{u\cdot X,v}/[\hat{V}(\hat{R}_{u\cdot X,v})]^{\frac{1}{2}} \tag{39}$$

can be used in a test of $H_0: R_{X,u+v} = R_{X,v}$ versus $H_1: R_{X,u+v} < R_{X,v}$ by comparing $z_{u\cdot X,v}$ with Z_α. This test is then used at each step of the stepwise, stepdown, and minimal best classification algorithms, analogously to those give for the case of no subsampling by Evans et al. (1985b), but with the following variations as proposed by Evans (1984).

In Step 0 of the stepwise algorithm here, enter the group g_1, along with X, that gives the maximum $z(g_1|X) > Z_\alpha$. In Step 1, enter as the (v+1)st group the group g_k, $k \in \{v+1,\cdots,T\}$, that gives the maximum $z(g_k|X,g_1,\cdots,g_v) > Z_\alpha$. In Step 2, drop the group $g_{k'}$, $k' \in \{1,\cdots,v\}$ that gives the minimum $z(g_{k'}|X,g_{1'},\cdots,g_{v'}) < Z_\alpha$, where $g_{1'},\cdots,g_{v'}$ are the other v of the v+1 currently entered groups.

To enable a simultaneous stepdown testing of the null hypothesis of no reduction in Bayes risk due to adding groups $g_1,\cdots,g_T$ to X, expand

$$\hat{R}(g_1,\cdots,g_T|X,g_0) \equiv \hat{R}(g_1,\cdots,g_T|X)$$
$$= \hat{R}(X) - \hat{R}(X,g_1,\cdots,g_T) \tag{40}$$

as

$$\hat{R}(g_1,\cdots,g_T|X) = \hat{R}(g_1|X) + \hat{R}(g_2|X,g_1)$$
$$+\cdots+ \hat{R}(g_T|X,g_1,\cdots,g_{T-1}) \quad . \tag{41}$$

Then each null hypothesis of no reduction in Bayes risk due to group g_k is tested by comparing $z(g_k|X,g_0,g_1,\cdots,g_{k-1})$ with Z_{α_k}, where $\alpha_k \simeq P[z(g_k|X,g_0,g_1,\cdots,g_{k-1}) > Z_{\alpha_k}]$. At the last step, Step T, if $z(g_1|X,g_0) \equiv z(g_1|X) \leq Z_{\alpha_1}$, then select no groups and retain only X for future allocations.

3.3. Variable Numbers of Subsamples Proportional to Sizes of Sampling Units — Misclassification Costs Proportional to Sizes

Based on $\bar{\underline{y}}_{\cdot}^{(v)}$ and X, the Bayes rules of Section 2.2.2 are used to classify the test units. Let $Q_{i'}^{(X,v)}$ be a variable that takes either value x is a test sampling unit is classified into class i' or 0 otherwise. To make $Q_{i'}^{(X,v)}$ relate specifically to class i, replace it by $Q_{i'\cdot i}^{(X,v)}$; then $Q_{i'\cdot i}^{(X,v)}$ takes either value $q_{i'ij}^{(X,v)} = x_{ij}$ if a test unit j from class i is classified into class i' or 0 otherwise. The Bayes risk formula of Equation (31) is used here with $Q_{i'}^{(X,v)}$ in place of $Q_{i'}$ to define $R_{X,v} = R(X,g_1,\cdots,g_v)$. Each term $E_i\left(Q_{i'}^{(X,v)}\right)$ is estimated and replaced by $\bar{q}_{i'i\cdot}^{(X,v)}$ to give $\hat{R}_{X,v}$. After replacing the definition of $Q_{i'\cdot i}^{(X,v)}$ of Section 3.2 by that above, Equation (37) is evaluated to give $z_{X,v}$ for use in an all-possible-subsets algorithm as given in Section 3.2. Similarly, the test statistic $z_{u\cdot X,v}$ given in Equation (39) can now be used to apply the stepwise, stepdown and minimal-best algorithms as given in Section 3.2.

4. A Remote Sensing Example

Background details for this example are given in Evans et al. (1985a). The quantities $\bar{\underline{y}}_{ij\cdot} = \left(\bar{y}'_{1ij\cdot},\cdots,\bar{y}'_{5ij\cdot}\right)'$, the number of subsampling units (=pixels) n_{ij}, and $\sum_{k=1}^{n_{ij}}(\underline{y}_{ijk} - \bar{\underline{y}}_{ij\cdot})(\underline{y}_{ijk} - \bar{\underline{y}}_{ij\cdot})'$ were available on each calibration sampling unit (=paddock or field) $j=1,\cdots,r_i$ from each class $i=1,\cdots,5$ ($r_1=9$, $r_2=8$, $r_3=4$, $r_4=7$, and $r_5=6$), but only the $\bar{\underline{y}}_{ij\cdot}$ were used in Evans et al. (1985a,b). Here, the other quantities are also used to

calculate $\bar{y}_{i..}$, S_{δ_i}, and S_i of Equations (4), (5), and (7), and thence S_δ, S_ε, S, $\hat{\Sigma}_\varepsilon$, and $\hat{\Sigma} = \hat{\Sigma}_\varepsilon + S_\delta/n$, for an arbitrarily chosen n=50, of Equations (16)–(20). Then $\bar{y}_{i..}$ and $\hat{\Sigma}$ are regarded as the parameters μ_i, i=1,···,5, and common $\Sigma = \Sigma_\varepsilon + \Sigma_\delta/n$, respectively, of five multivariate normal distributions from each of which to simulate r_i=10 new calibration observations and m_i=10 test observations, $\bar{y}_{ij.}$, for j=1,···,20, i=1,···,5, each based on a constant n=50 subsamples for the sake of simplicity. To cover a range of differences among classes, the μ_i were modified by adding and subtracting constants to give two extra sets of class mean vectors and Σ was redefined with n=20, thus giving a total of six different sets of distributions from each of which was simulated a data set.

For each simulated data set, the N=50 calibration observations were used to calculate $\bar{y}_{i..}$, i=1,···,5, and $\hat{\Sigma}$ of Equation (21), appropriate subvectors $\bar{y}^{(v)}_{i..}$ and submatrices $\hat{\Sigma}^{(v)}$, which were then used in the Bayes rule of Section 2.1 to classify the M=50 test observations $\bar{y}^{(v)}_{ij.}$ according to each subset of v=1,···,5 groups. A totally random allocation, i.e., using no groups, was also made of the M test observations. Equal prior probabilities π_i=1/5 and unequal misclassification costs were used in each Bayes rule and in the estimation of Bayes risks and their variances and covariances.

The MATRIX procedure of the SAS package (SAS Institute Inc., 1982) was used to implement the methods above as well as stepwise, stepdown (for chronological and reverse testing order as in Evans et al. (1985b)), and minimal-best algorithms of Section 3.1, for each data set. Either the original programs can be found in Evans (1984) or updates obtained from the first author of this paper. A significance level of α=0.05 was used at

each step of the stepwise and minimal-best algorithms and overall for the stepdown algorithm. For both the stepdown reverse chronological and minimal-best algorithms, the selected subset of groups was near optimal for four out of six data sets, although for two of those in the stepdown case all groups were retained and thus no dimension reduction was achieved. In contrast, the stepwise and chronological stepdown algorithms chose zero ad one subsets, respectively, that were optimal; in the one optimal subset, all groups were retained. For each algorithm, every suboptimal subset arose due to too few groups being retained, even though for example in the stepwise case, all six subsets were the best of the same size. Of the suboptimal selected subsets, their z-values compared to those of the best subset of any size 1, 2, 3, or 4 were inflated by 110% and 50% in the case of minimal best; 200%, 122%, 111%, 222%, 210%, and 50% for stepwise; 260%, 144%, 460%, 210%, and 50% for chronological stepdown; and 253% in the case of reverse stepdown.

Based on these data sets, the minimal-best classification algorithm is superior to the others implemented here. But it is clear that variations of these algorithms should be considered in future studies. In particular, if computing expense is not prohibitive, variations of the minimal-best algorithm, as suggested by Evans et al. (1985b), should provide optimal subsets in most cases. If computing is too prohibitive for studying all-possible-subsets, then improved variations of stepping algorithms need to be developed.

5. References

Arnold, S. F. (1981), *The Theory of Linear Models and Multivariate Analysis* (Wiley-Interscience, New York).

Evans, J. C. (1984). Stagewise Selection and Classification of Multivariate Repeated Measurements, Ph.D. Thesis (Cornell University, 1983). (University Microfilms International, Ann Arbor, Michigan).

Evans, J. C., Robson, D. S., and Schwager, S. J. (1985a), The use of estimated Bayes risk as a criterion for selecting subsets of allocation variables.

Evans, J. C., Robson, D. S., and Schwager, S. J. (1985b), Stagewise classification algorithms for selecting a subset of groups of allocation variables.

Harville, D. A. (1977), Maximum likelihood approaches to variance component estimation and to related problems, *J. Amer. Statist. Assoc.* 72, 320-337.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974), *Introduction to the Theory of Statistics* (McGraw-Hill, New York, 3rd ed.).

SAS Institute Inc. (1982), *SAS Users' Guide: Statistics,* 1982 Edition (Cary, North Carolina).

Searle, S. R. (1971), *Linear Models* (Wiley, New York).

nep